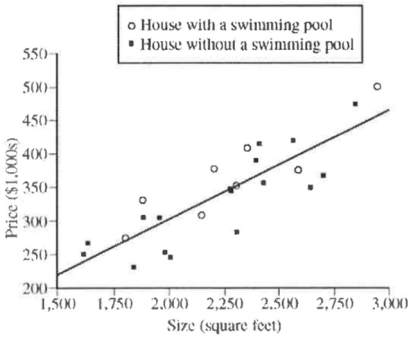
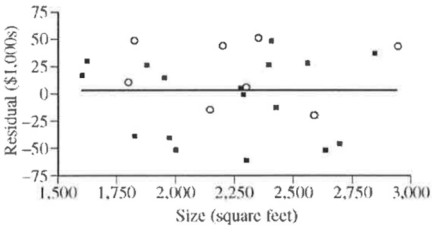
Name:_____

AP Stats Assignment 15.4 Confidence Interval for Slope of LSRL

1. A real estate agent is interested in developing a model to estimate the prices of houses in particular part of a large city. She takes a random sample of 25 recent sales and, for each house, records the price (in thousands of dollars), the size of the house (in square feet), and whether or not the house has a swimming pool. This information, along with regression output for a linear model using size to predict price, is shown below and on the next page.

Price ($1,000s)	Size (square feet)	Pool	Residual ($1,000s)
274	1,799	yes	6
330	1,875	yes	49
307	2,145	yes	–18
376	2,200	yes	42
352	2,300	yes	1
409	2,350	yes	50
375	2,589	yes	–23
498	2,943	yes	42
248	1,600	no	13
265	1,623	no	26
228	1,829	no	–45
303	1,875	no	22
303	1,950	no	10
251	1,975	no	–46
244	2,000	no	–57
347	2,274	no	1
345	2,279	no	–2
282	2,300	no	–69
389	2,392	no	23
413	2,410	no	44
353	2,428	no	–19
419	2,560	no	26
348	2,639	no	–58
365	2,701	no	–52
474	2,849	no	33





Linear Fit
Price = –28.144 + 0.165 Size

Summary of Fit
RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

a) Interpret the slope of the LSRL in the context of the study

b) Create a 95% confidence interval for the slope of the LSRL

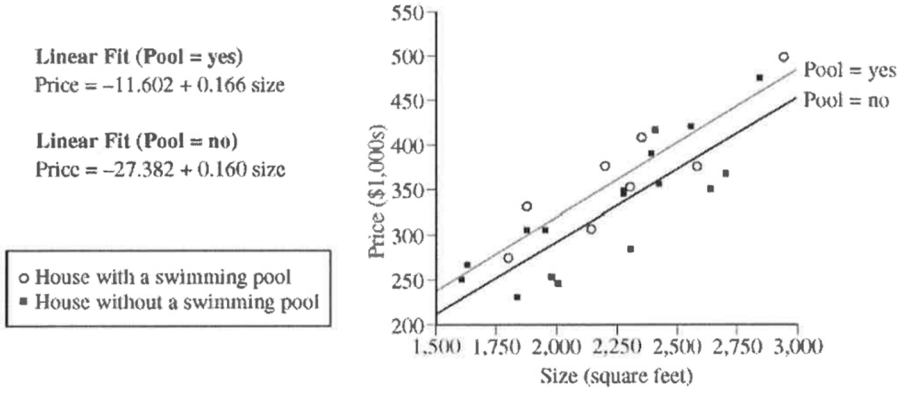
c) The third house in the table has a residual of 18. Interpret this residual value in the context of the study.

d) The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house. Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool

e) To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models in shown below. The conditions for inference have been checked and verified, and a 95 percentage confidence interval for the true difference in the two slopes is (-0.099, 0.110). Based on this interval, is there a significance difference in the two slopes? Explain:

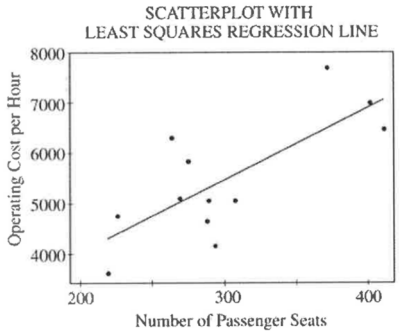
Linear Fit (Pool = yes)
Price = –11.602 + 0.166 size

Linear Fit (Pool = no)
Price = –27.382 + 0.160 size



f) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part d.

2. Commercial airlines need to know the operating cost per hour of flight for each plane in their fleet. In a study of the relationship between operating cost per hour and number of passenger seats, investigators computed the regression of operating cost per hour on the number of passenger seats. The 12 sample aircraft used in the study included planes with as few as 216 passenger seats and planes with as many as 410 passenger seats. Operating cost per hour ranged between $3600 and $7800. Some computer output from a regression analysis of these data is shown below.



Predictor	Coef	StDev	T	P
Constant	1136	1226	0.93	0.376
Seats	14.673	4.027	3.64	0.005
S = 845.3	R-Sq = 57.0%		R-Sq (adj) = 52.7%	

a) What is the equation of the LSRL regression line that describes the relationship between operating cost per hour and the number of passenger seats in the plane? Define any variables used in this equation

b) What is the predicted cost per hour for a plane with 350 seats?

c) What is the value of the correlation coefficient for operating cost per hour and the number of passenger seats in the plane? Interpret this correlation.

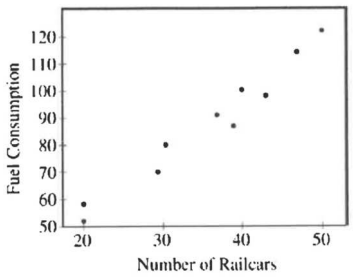
d) Create a 99% confidence interval for the slope of the LSRL. Interpret this slope.

e) The standard deviation of the regression line is S=845.3. Interpret this value.

f) Suppose that you want to describe the relationship between operating cost per hour and the number of passenger seats in the plane for planes only in the range of 250 to 350 seats. Does this line shown in the scatterplot still provide the best description of the relationship for data in this range? Explain why or why not.

3. The Great Plains Railroad is interpreted in studying how fuel consumption is related to the number of railcars for its trains on a certain route between Oklahoma City and Omaha. A random sample of 10 trains on this route has yielded the data shown in the table below. A scatterplot, residual plot, and the output from the regression analysis for these data are shown below.

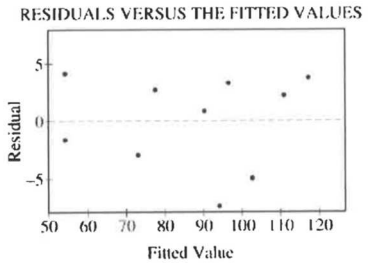
Number of Railcars	Fuel Consumption (units/mile)
20	58
20	52
37	91
31	80
47	114
43	98
39	87
50	122
40	100
29	70



The regression equation is
Fuel Consumption = 10.7 + 2.15 Railcars

Predictor	Coef	StDev	T	P
Constant	10.677	5.157	2.07	0.072
Railcar	2.1495	0.1396	15.40	0.000

S = 4.361 R-Sq = 96.7% R-Sq(adj) = 96.3%



a) In a Linear model appropriate for modeling these data? Clearly explain your reasoning

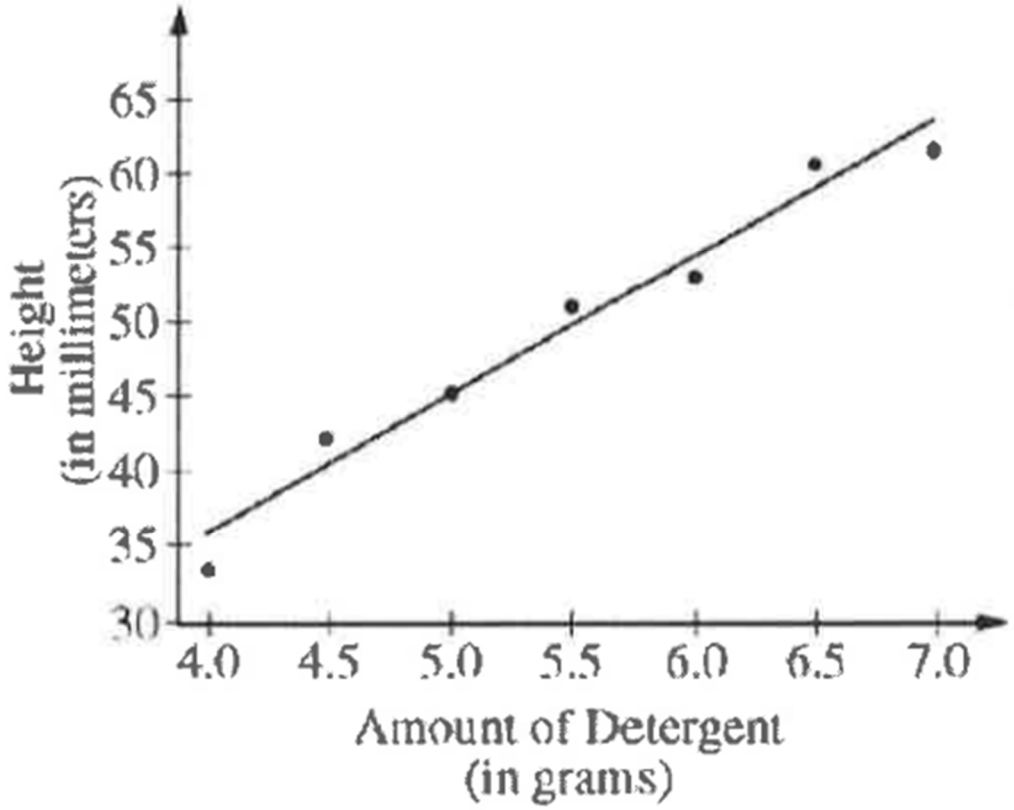
b) Suppose the fuel consumption cost is $25 per unit. Give a point estimate (single value) for the change in the average cost of fuel per mile for each additional railcar attached to a train. Show your work

c) interpret the value of r^2 in the context of this problem

d) Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train has 65 railcars? Explain:

A manufacturer of dish detergent believes the height of soapsuds in the dishpan depends on the amount of detergent used. A study of the suds' heights for a new dish detergent was conducted. Seven pans of water were prepared. All pans were of the same size and type and contained the same amount of water. The temperature of the water was the same for each pan. An amount of dish detergent was assigned at random to each pan, and that amount of detergent was added to the pan. Then the water in the dishpan was agitated for a set amount of time, and the height of the resulting suds was measured.

A plot of the data and the computer output from fitting a least squares regression line to the data are shown below.



Predictor	Coef	SE Coef	T	P
Constant	−2.679	4.222	−0.63	0.554
Amount	9.5000	0.7553	12.58	0.000

S = 1.99821 R–Sq = 96.9% R–Sq(adj) = 96.3%

(a) Write the equation of the fitted regression line. Define any variables used in this equation.

(b) Note that $s = 1.99821$ in the computer output. Interpret this value in the context of this study.

(c) Identify and interpret the standard error of the slope.

6. A study was designed to explore subjects' ability to judge the distance between two objects placed in a dimly lit room. The researcher suspected that the subjects would generally overestimate the distance between the objects in the room and that this overestimation would increase the farther apart the objects were.

   The two objects were placed at random locations in the room before a subject estimated the distance (in feet) between those two objects. After each subject estimated the distance, the locations of the objects were rerandomized before the next subject viewed the room.

   After data were collected for 40 subjects, two linear models were fit in an attempt to describe the relationship between the subjects' perceived distances (y) and the actual distance, in feet, between the two objects.

   Model 1: $\hat{y} = 0.238 + 1.080 \times$ (actual distance)

   The standard errors of the estimated coefficients for Model 1 are 0.260 and 0.118, respectively.

   Model 2: $\hat{y} = 1.102 \times$ (actual distance)

   The standard error of the estimated coefficient for Model 2 is 0.393.

(a) Provide an interpretation in context for the estimated slope in Model 1.

(b) Explain why the researcher might prefer Model 2 to Model 1 in this context.

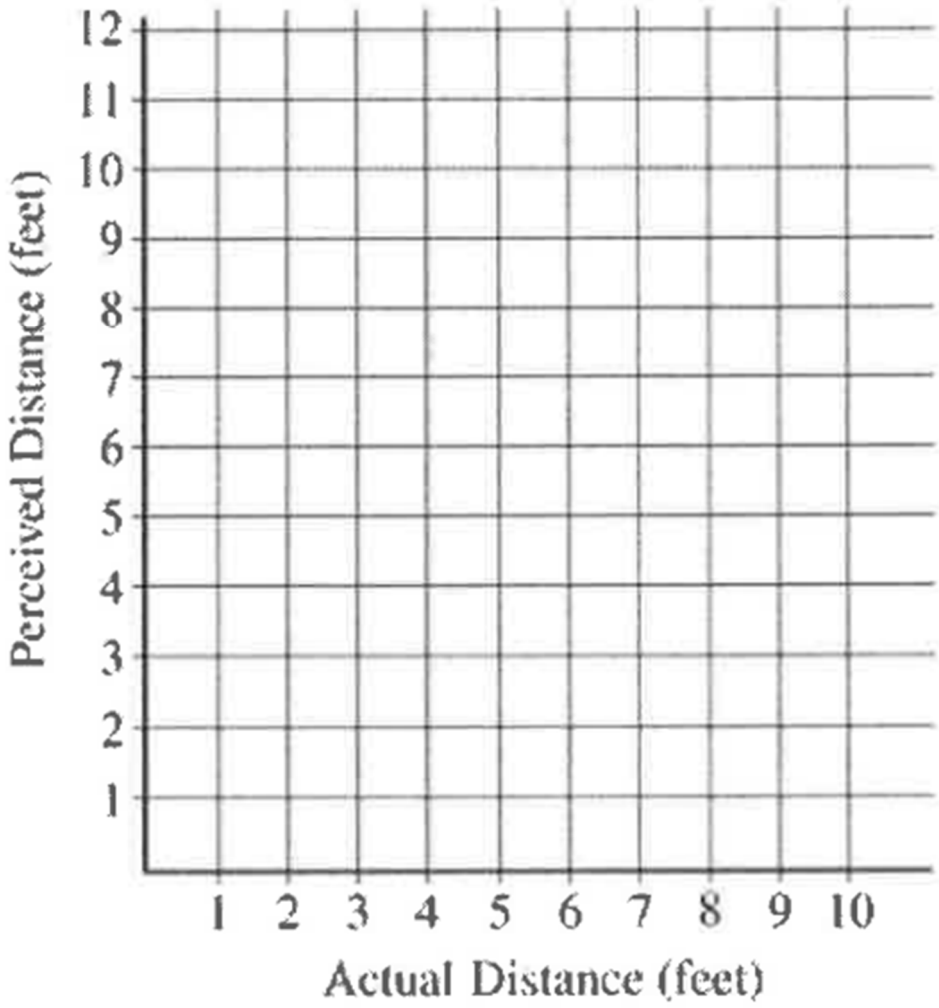
(c) Using Model 2, test the researcher's hypothesis that in dim light participants overestimate the distance, with the overestimate increasing as the actual distance increases. (Assume appropriate conditions for inference are met.)

The researchers also wanted to explore whether the performance on this task differed between subjects who wear contact lenses and subjects who do not wear contact lenses. A new variable was created to indicate whether or not a subject wears contact lenses. The data for this variable were coded numerically (1 = contact wearer, 0 = noncontact wearer), and this new variable, named "contact," was included in the following model.

Model 3: $\hat{y} = 1.05 \times (\text{actual distance}) + 0.12 \times (\text{contact}) \times (\text{actual distance})$

The standard errors of the estimated coefficients for Model 3 are 0.357 and 0.032, respectively.

(d) Using Model 3, sketch the estimated regression model for contact wearers and the estimated regression model for noncontact wearers on the grid below.



(e) In the context of this study, provide an interpretation of the estimated coefficients for Model 3.